

# Multimodal Automatic Image Annotation Method using Association Rules Mining and Clustering

Mounira Taileb, Eman Alahmadi  
Faculty of Computing and Information Technology  
King Abdulaziz University  
Saudi Arabia, Jeddah

**Abstract**—Effective and fast retrieval of images from image datasets is not an easy task, especially with the continuous and fast growth of digital images added everyday by used to the web. Automatic image annotation is an approach that has been proposed to facilitate the retrieval of images semantically related to a query image. A multimodal image annotation method is proposed in this paper. The goal is to benefit from the visual features extracted from images and their associated user tags. The proposed method relies on clustering to regroup the text and visual features into clusters and on association rules mining to generate the rules that associate text clusters to visual clusters. In the experimental evaluation, two datasets of the photo annotation tasks are considered; ImageCLEF 2011 and ImageCLEF 2012. Results achieved by the proposed method are better than all the multimodal methods of participants in ImageCLEF 2011 photo annotation task and state-of-the-art methods. Moreover, the MiAP of the proposed method is better than the MiAP of 7 participants out of 11 when using ImageCLEF 2012 in the evaluation.

**Keywords**—Automatic image annotation; association rules mining; clustering

## I. INTRODUCTION

Nowadays, we are witnessing an enormous increase in the number of images available on the web which makes image retrieval (IR) a challenging task. In literature, content-based image retrieval (CBIR) and text-based image retrieval (TBIR) are the main two approaches to achieve the IR. In the TBIR systems, the retrieval of images relies on the text or keywords (called also labels or concepts) entered by users; such systems depend mainly on the image tags typed manually by human and/or the text accompanying and describing the image on the Web page. The main disadvantages of TBIR systems are: (1) the inaccuracy of the manual annotation, (2) manual annotation is not always available and impossible for large image database. Whereas, in the CBIR systems, the retrieval relies only on the visual content of images which is represented by visual features such as texture, edge, color, etc. CBIR systems suffer from a well known problem called the semantic gap problem [1], [2] which refers to the difference between the low level visual features of the image and the high-level user semantic concept. The automatic image annotation (AIA) is a way to address this issue; it has received an increasing attention from researchers and has become an important field of research for image retrieval [3], [4], [5]. Indeed, AIA facilitates the search in huge datasets of images and improve the retrieval of images that are semantically similar to a query image. AIA is the process of automatically assigning keywords from a

predefined vocabulary to an image, keywords that characterize its semantic visual content. Several approaches exist in the AIA, they are classified into three categories [6] as illustrated in Fig. 1.

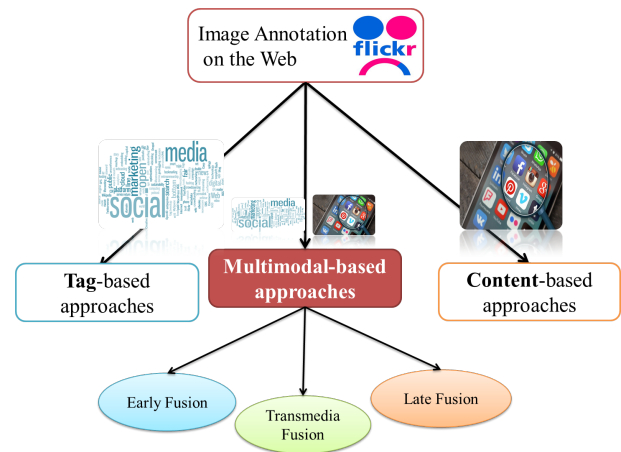


Fig. 1. Automatic image annotation approaches

In tag-based image annotation [7], [8], the tags used to annotate an image are collected from the predefined user tags and the articles around the image. However, in content-based image annotation, only the visual features are used [9], [10]. And in multimodal-based approaches both modalities (tag and visual features) are utilized to annotate an image [6], [11], [12]. In this paper, a new multimodal image annotation method is proposed using association rules mining (ARM) and clustering techniques. The ARM is used to explore the semantic relations that might exist between text clusters and visual feature clusters in order to associate them. The goal is to combine visual and text modalities to improve the image annotation performance. The proposed method consists of two phases; the training phase and the annotation phase. In the training phase, the input is a set of tagged or labeled images. The visual features of images are clustered as well as the tags of images which are considered as text features. Then the ARM is used to fuse the visual clusters and text clusters and find relations between them. The training phase provides at the end a list of association rules. In the annotation phase, a new image is annotated using its visual features which are used to extract the related association rules to provide a list of tags to the image.

The remaining part of the paper is structured as follows.

In Section 2, the different approaches in AIA are presented. Then Section 3 describes the association rules mining. Section 4 provides a detailed description of the proposed method with its training and testing phases. The experimental evaluation is detailed in Section 5. The paper conclusion is provided in section 6.

## II. RELATED WORK

As mentioned above, several approaches can be used to solve the problem of AIA, they are categorized into three categories: tag-based approaches, content-based approaches and multimodal-based approaches. In the tag-based approach, images are retrieved and annotated based on the text or keywords entered by users; such systems depend mainly on the image tags typed manually by humans and/or the text provided with the image on the web page. The goal of the methods proposed in the tag-based approach is whether the refinement or enrichment of the image annotation. In some methods [8], [13] a list of tags is refined and reduced to a smaller list of reliable tags. And in other methods [6], [7], [12], [14] a short list of tags are enriched after performing AIA.

In content-based annotation approach, several methods have been proposed [5], [10], [15], [16], [17], [18]. In this approach, a set of training images is used with the image annotation keywords. The images are segmented into regions [19] to create a high-quality segmentation, then the visual features of regions are extracted and clustered to obtain the blobs. A blob refers to the label associated to a region. After that, blobs are linked with the keywords Fig. 2, and this is the key process of the content-based annotation approach. The linking process can be performed using several methods, such as: Expected Maximum (EM) in the translation model [15], the latent Dirichlet allocation (LDA) model [16], the co-occurrence model [17], the Cross-Media Relevance Model (CMRM) [10] and Multiple Bernoulli Relevance Model (MBRM) [5]. To be annotated, a new image is segmented to extract regions, and the features of each region are utilized to determine its corresponding blob. Finally, keywords are predicted from the corresponding blobs to annotate the image. The methods proposed in multimodal-based annotation approach focus on leveraging both modalities (tag-based and content-based) to improve the automatic image annotation performance. Many studies have been conducted in this area, falling into three categories: the early fusion, transmedia fusion and late fusion [20], [21], [22]. The early fusion consists of combining the visual and text features into a single representation before the annotation process. In the transmedia fusion, one of the two modalities (visual or text) is used to gather the relevant images before switching to the other modality to aggregate tag features of these relevant images and perform the image annotation. However, in the late fusion, there is a separation in each modality processing, then the results of two previous processes are combined at the decision level of the annotation process.

## III. ASSOCIATION RULES MINING

One of the well-known and widely used data-mining technique is the association rules mining. In a knowledge discovery operation, data mining is defined as a procedure that attempts to discover a new and meaningful pattern in data. Analyzing the association rules between groups of items in a set is useful

for uncovering the interesting relationships that may be hidden in a huge database. The example of the content of a market baskets is considered a classical example for the extraction of association rules. Items are defined as things that anyone may buy from the market, and transactions are the diverse items contained in market baskets. Below is an example of a simple association rule extracted from market baskets:

Bread  $\rightarrow$  Eggs

This rule shows that there is a strong relationship between the selling of bread and eggs, which results when customers who buy bread also buy eggs. The goal of discovering such rules is to describe the customers purchase behavior, which can help companies to find opportunities for sales and better guide category management in order to increase profits. Association-rule mining is usable in several applications in diverse fields, such as web mining, Bioinformatics, and medical diagnosis. The method proposed in this paper aims to provide a semantic annotation by using the ARM algorithm to perform an association between text clusters and visual clusters that are semantically related. Generating a good set of ARs depends mainly on the setup of support and confidence, their calculations are given in the equations (1) and (2) respectively as defined in [23]:

$$Supp(X) = \frac{count(X)}{N} \quad (1)$$

Where  $N$  refers to all transactions in the transaction database  $T$ .

$$Conf(X \rightarrow Y) = \frac{count(X \cup Y)}{count(X)} \quad (2)$$

In the rule  $X \rightarrow Y$ , the calculation that defines how many items in  $Y$  appear also in other transactions that include  $X$  is known as confidence of the rule. While determining how often a rule is applied in a given database is known as the support of the rule.

## IV. PROPOSED METHOD

A novel multimodal image annotation method using association rules mining and clustering algorithms is proposed, it takes advantage of using both text and visual modalities to improve AIA. The proposed method uses clustering algorithms to regroup text and visual features into clusters, then association rules mining are applied to generate the rules that associate text clusters to visual clusters. This can be considered as a late fusion between text and visual clusters. The novelty of the proposed method is in the use of clustering and associations rules mining. The Fig. 2 is an illustration of the method framework. The method comprises two main phases: the training phase and the annotation phase. In the first phase, the training dataset includes images with their associated text files (each image has a text file containing the list of user tags) [14]. All the steps of the training phase are described in the following sub-sections.

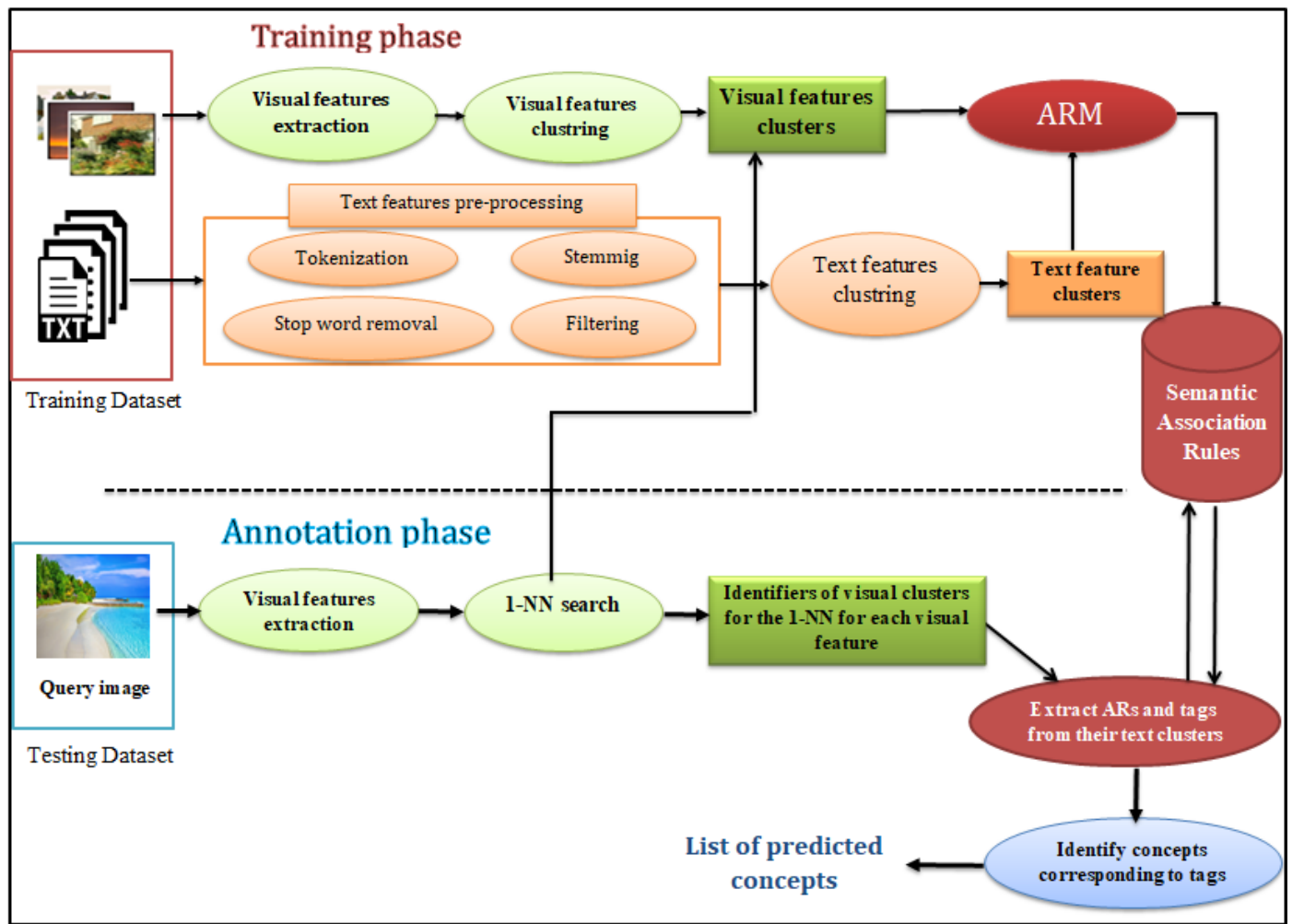


Fig. 2. Flowchart of the proposed annotation method

### A. Training Phase

In the training phase, a series of processes is performed. It starts by extraction of visual and text features, then the clustering of both modalities. And ends with the generation of association rules. The different processes are detailed in the following.

1) *Features Extraction:* In this phase, two MPEG-7 global visual features are extracted; CSD (color structure descriptor) and EHD (edge histogram descriptor). In the other hand, the text files associated to images are considered as text features, they are first pre-processed by performing tokenization, stemming, stop word removal and filtering. The latter pre-processing step is performed by comparing the set of all words with American and British English dictionaries to ignore any non-English words. For example, the number of words (tags) in the dataset ImageCLEF 2011 was 88,083 and after pre-processing, 52,527 words remained. The number of remaining words have been reduced to less than 52,527 words, but we have noticed that short words like countries and cities names were removed (such as USA), hence, the final number of words was maintained at 52,527 words. It is worth to mention that only a few of the remaining words have been

manually associated to the concepts predefined in the two datasets (ImageCLEF 2011 and ImageCLEF 2012) used in the experimental evaluation of the proposed annotation method. As an example, from ImageCLEF 2011, the following tags: cello, bass, viola, guitar, piano, drum, and organ are associated to the concept called Musical Instrument. For the dataset ImageCLEF 2011, the total number of words associated to concepts is 292, but these words are repeated several times in the final list of filtered words.

2) *Clustering:* The features extracted in the previous step are regrouped into clusters using clustering algorithms. The visual features are clustered using the hierarchical clustering method, also called indexing technique, NOHIS [24]. This indexing technique groups the visual features into clusters and the clusters are organized in a hierarchal structure called NOHIS-tree. For each of the visual features (CSD and EHD), a NOHIS-tree is constructed. The text features are clustered as well using K-means [25]. After the clustering of both visual and text features, the ARM algorithm is performed on visual and text clusters to find relations between them. The output is the list of association rules that will be used later in the annotation phase.

3) *Generation of Association Rules:* In order to generate the association rules, it is important to define first set of items (itemsets)  $I$  and the transaction database  $T$ . In the proposed method, the text and visual clusters represent the itemsets, where the text clusters are denoted by  $C_{ti}$  and the visual features clusters are denoted by  $C_{cj}$  for color clusters and  $C_{ek}$  for edge clusters. After determining the features model space for all of the feature modalities, the transaction database  $T$  can be constructed and the ARM algorithm can be run over it.

The relationship between one text cluster  $C_{ti}$  and at least one of the visual clusters (color or/and edge,  $C_c$  and  $C_e$  respectively) is considered as transaction. In other words, the association is between a text cluster and one or two visual clusters is a transaction. A text cluster is associated to one or two visual clusters if the clusters of different modalities have images in common (since each tag in the text clusters and each visual feature in the visual clusters belong to an image). Hence, a transaction is constructed if the number of common images that are in both text and visual clusters is greater than zero, as illustrated in the following example:

If  $|C_{ti} \cap C_{cj}| > 0$ , then  $\{C_{ti}, C_{cj}\}$  is added to  $T$

Examples of transactions are given in the following:

$\{C_{t0}, C_{c62}\}, \{C_{t33}, C_{e57}\}, \{C_{t46}, C_{c25}\},$   
 $\{C_{t51}, C_{c42}, C_{e75}\}, \{C_{t68}, C_{c67}, C_{e76}\}.$

The formal equations of support and confidence have been modified as in [26] to be adapted to the proposed method for the following reasons; first, if the calculated support/confidence values of the association rules for the whole transaction database  $T$  result in a low support value, this could affect later the generated association rules. Second, the goal is to explore the semantic relations among text and visual clusters, hence the support and confidence calculation have to be restricted to the text clusters results. Therefore, the support and confidence of the rule defined as  $C_{ti} \rightarrow C_{vj}$  (where  $C_{vj}$  refers to the visual cluster) are as follows:

$$Supp(C_{ti} \rightarrow C_{vj}) = \frac{count(C_{ti}, C_{vj})}{count(C_{ti})} \quad (3)$$

$$Conf(C_{ti} \rightarrow C_{vj}) = \frac{count(C_{ti}, C_{vj})}{max_k(count(C_{ti}, C_{vk}))} \quad (4)$$

When there are multiple items on the right side, the similarity of the rule is as follows:

$$Supp(C_{ti} \rightarrow C_{vj} | j = 1, \dots, m) = \frac{count(C_{ti}, C_{vj} | j = 1, \dots, m)}{count(C_{ti})} \quad (5)$$

$$Conf(C_{ti} \rightarrow C_{vj} | j = 1, \dots, m) = \frac{count(C_{ti}, C_{vj} | j = 1, \dots, m)}{max_k(count(C_{ti}, C_{vk}))} \quad (6)$$

The modified definitions of support and confidence are used to identify the frequent itemsets by applying Apriori algorithm [27] as defined in [26]. The algorithm needs as parameters the minimum value of support  $minsup$  and the transaction database  $T$ . Then the frequent itemsets are used along with the minimum value of confidence  $minconf$  to generate the association rules.

Each association rule is stored with its support and confidence values, as shown in Fig. 3. Later, these association rules will be retrieved during the annotation phase to predict the list of tags (or concepts) to annotate a new image.

$\{C_{t2}\} \Rightarrow \{C_{c82}\}$  (support: 6.9 confidence: 80 )  
 $\{C_{t22}\} \Rightarrow \{C_{c66}\}$  (support: 33.33 confidence: 100 )  
 $\{C_{t31}\} \Rightarrow \{C_{e73}\}$  (support: 2.48 confidence: 91.3 )  
 $\{C_{t74}\} \Rightarrow \{C_{c85}\}$  (support: 66.67 confidence: 100 )

Fig. 3. Example of generated associations rules

It is worth to mention that the association rules has been used in a tag-based image annotation method [13], in this method the association rules is employed in a completely different way where each web page is considered as a transaction and the set of words in the web page is regarded as the itemset.

### B. Annotation Phase

In the annotation phase, and for each image of the test dataset, the following processes are performed:

- Extraction of the color and edge visual features (CSD and EHD respectively).
- Search of the first nearest neighbor 1-NN for the visual features CSD and EHD extracted in the previous step. The search is performed on the two hierarchical structures NOHIS-tree constructed in the training phase for the two visual feature spaces. The distance used in the nearest neighbor search is the Euclidian distance. The nearest neighbor for each visual feature is returned along with its cluster identifier. In other words, the identifier of the visual cluster that contains the nearest neighbor of the visual feature CSD (extracted from the test image) is returned. And the same is done for the visual feature EHD. An example is provided in Fig. 4, where  $C_{c66}$  and  $C_{e54}$  are the clusters that contain the nearest neighbor of CSD and EHD features respectively.
- Each of the visual cluster identifiers is used to extract all the rules that contain this visual cluster. The text clusters of all the extracted rules are considered by taking the tags they contain, as shown in Fig. 4.
- The concepts that correspond to the previous tags are considered as annotation of the image. We remind that the image datasets used in the experimental results are using a list of concepts to annotate the image, and the tags have been linked to the concepts in the training phase.

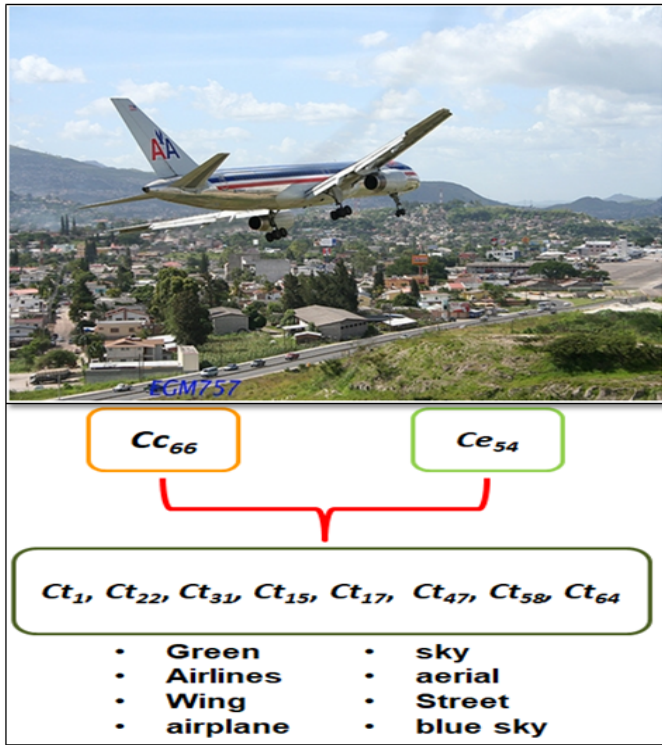


Fig. 4. List of tags extracted from the text clusters

## V. EXPERIMENTAL EVALUATION

Two datasets are used in the performance evaluation of the proposed automatic annotation method; ImageCLEF 2011 and ImageCLEF 2012. These datasets are considered because of two reasons; the availability of ground truth and their wide use. In the following, a detailed description is provided regarding the datasets, evaluation metrics and results. The values of *minsupp* and *minconf* are set respectively at 2% and 70% based on a set of experiments.

### A. Datasets

The dataset of the annotation task in ImageCLEF 2011<sup>1</sup> is used; it consists of a training dataset of 8000 images with their associated user tags (in text files), and a test dataset of 10,000 images. The objective is to annotate the test dataset with 99 concepts. The second dataset is ImageCLEF 2012<sup>2</sup> used in the sub-task 1; concept annotation, for the year 2012. The size of the training dataset is 15,000 images provided with used tags and the test dataset is a collection of 10,000 images to annotate with 94 concepts.

### B. Evaluation metrics

The mean interpolated average precision MiAP is used for the evaluation per concept, and F-Measure is used for the evaluation per example (per image). To calculate the MiAP, first 11 recall values are considered from 0.0 to 1.0 with steps

of 0.1. Then for each recall value  $R$ , its interpolated precision  $P_{interp}$  is the highest precision of any recall value  $R' \geq R$ :

$$P_{interp}(R) = \max_{R' \geq R} P(R) \quad (7)$$

Then the average interpolated precision of the 11 recall values is calculated as follows:

$$AP_{interp} = \frac{1}{11} \sum_{R=0}^1 P_{interp}(R) \quad (8)$$

The MiAP is the average of the average interpolated precisions of all concepts, it is calculated as follows:

$$MiAP = \frac{1}{C} \sum_{i=0}^c AP_{interp} \quad (9)$$

$C$  is the number of concepts used for the annotation.

The calculation of F-Measure is given in the following equation:

$$F - Measure = \frac{2(P * R)}{(P + R)} \quad (10)$$

Where  $P$  is the precision and  $R$  is the recall for each annotated image.

### C. Results

The results obtained with the proposed method are compared to results of multimodal methods proposed by participants in the annotation tasks of ImageCLEF 2011 and ImageCLEF 2012 and methods with best results we found in literature [6], [12], [21]. The MiAP<sup>3</sup> and F-Measure<sup>4</sup> are used in the comparison with participants. However, only the MiAP is available for the state-of-the-art methods.

The proposed method achieved results that outperforms those of all methods considered in table I in terms of MiAP and F-Measure. The proposed method obtains a gain of 23.3% of MiAP compared to the best MiAP achieved by the participant "TUBFI", and a gain of 17.7% of F-Measure compared to the best F-Measure achieved by the participant "ISIS". In addition, the MiAP obtained with the proposed method outperforms the best MiAP of [12] by 22.8% as show in in table II.

The table III illustrates the comparison of the proposed method results to results (of the multimodal methods) of participants in ImageCLEF 2012 photo annotation task and the multimodal annotation method proposed in [6]. The proposed method outperforms the results of 7 participants out of 11. The best MiAP outperforms ours by 11%. This is due to the few number of tags that have associated to the predefined list of concepts.

An example of images annotated with the proposed method is illustrated in Fig. 5. The concepts predicted by the proposed method are provided along with ground truth concepts by indicating the correct and wrong predicted labels.

<sup>1</sup><https://www.imageclef.org/2011/Photo>

<sup>2</sup><https://www.imageclef.org/2012/photo-flickr>

<sup>3</sup><https://www.imageclef.org/2011/PhotoAnnotationMAPResults>

<sup>4</sup><https://www.imageclef.org/2011/PhotoAnnotationExampleBasedResults>




Image	Proposed method concepts	Ground truth concepts
	<ul style="list-style-type: none"> <li>✓ Building_Sights</li> <li>✓ Citylife</li> <li>✓ Summer</li> <li>✓ Outdoor</li> <li>✓ Plants</li> <li>✓ Trees</li> <li>✓ Sky</li> <li>✓ Day</li> <li>✓ Neutral_Illumination</li> <li>✓ No_Blur</li> <li>✓ No_Persons</li> <li>✓ Architecture</li> <li>✓ Visual_Arts</li> <li>✓ natural</li> <li>✓ cute</li> </ul>	<ul style="list-style-type: none"> <li>Building_Sights</li> <li>Citylife</li> <li>Summer</li> <li>Outdoor</li> <li>Plants</li> <li>Trees</li> <li>Sky</li> <li>Day</li> <li>Neutral_Illumination</li> <li>No_Blur</li> <li>No_Persons</li> <li>Architecture</li> <li>Visual_Arts</li> <li>natural</li> <li>cute</li> </ul>
	<ul style="list-style-type: none"> <li>✓ Outdoor</li> <li>✓ Sky</li> <li>✓ Clouds</li> <li>✓ Water</li> <li>✓ Lake</li> <li>✓ Mountains</li> <li>✓ Neutral_Illumination</li> <li>✓ No_Blur</li> <li>✓ Shadow</li> <li>✓ Painting</li> <li>✓ cute</li> <li>✗ technical</li> </ul>	<ul style="list-style-type: none"> <li>Landscape_Nature</li> <li>Outdoor</li> <li>Sky</li> <li>Clouds</li> <li>Water</li> <li>Lake</li> <li>Mountains</li> <li>Day</li> <li>Neutral_Illumination</li> <li>No_Blur</li> <li>No_Persons</li> <li>Aesthetic_Impression</li> <li>Fancy</li> <li>Shadow</li> <li>Painting</li> <li>cute</li> <li>calm</li> </ul>
	<ul style="list-style-type: none"> <li>✓ Landscape_Nature</li> <li>✓ Summer</li> <li>✓ Outdoor</li> <li>✓ Plants</li> <li>✓ Flowers</li> <li>✓ Sky</li> <li>✓ Day</li> <li>✓ No_Persons</li> <li>✓ Aesthetic_Impression</li> <li>✓ Park_Garden</li> <li>✓ Visual_Arts</li> <li>✓ natural</li> <li>✗ Still_Life</li> </ul>	<ul style="list-style-type: none"> <li>Landscape_Nature</li> <li>Summer</li> <li>Outdoor</li> <li>Plants</li> <li>Flowers</li> <li>Sky</li> <li>Day</li> <li>Clouds</li> <li>Day</li> <li>Sunny</li> <li>Partly_Blurred</li> <li>No_Persons</li> <li>Aesthetic_Impression</li> <li>Park_Garden</li> <li>Visual_Arts</li> <li>natural</li> <li>cute</li> <li>calm</li> </ul>

Fig. 5. Example of images annotated with the proposed method. The predicted concepts are compared to the ground truth concepts

TABLE I. RESULTS OF THE PROPOSED METHOD AND PARTICIPANTS IN IMAGECLEF 2011 IN TERMS OF MIAP AND F-MEASURE

Method	MiAP	F-Measure
MUFIN	0.299001	0.461820
MRIM	0.377179	0.552276
BPACAD	0.436294	0.593088
IDMT	0.370975	0.551224
ISIS	0.432758	0.622038
LIRIS	0.436968	0.566935
MLKD	0.401642	0.558795
TUBFI	0.443449	0.565980
<b>Our method</b>	<b>0.676530</b>	<b>0.799936</b>

TABLE II. RESULTS OF THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS FOR IMAGECLEF 2011 IN TERMS OF MIAP

Method	MiAP
Multimodal method proposed in [21]	43.6 %
Multimodal method proposed in [12]	45.3 %
Multimodal method proposed in [6]	44.8 %
<b>Our method</b>	<b>67.6 %</b>

TABLE III. COMPARISON OF THE PROPOSED METHOD TO RESULTS OF PARTICIPANTS IN IMAGECLEF 2012 PHOTO ANNOTATION TASK IN TERMS OF MIAP

Participants in ImageCLEF 2012 photo annotation task	Method	MiAP
	BUA A AUDR	13.07 %
	CEA LIST	40.86 %
	CERTH	32.10 %
	DMS-SZTAKI	42.58 %
	ISI	41.36 %
	KIDS NUTN	17.17 %
	LIRIS	<b>43.67 %</b>
	MLKD	31.85 %
	UAIC	18.63 %
UNED	7.56 %	
URJCyUNED	6.22 %	
State-of-the-art multimodal method	Multimodal method proposed in [6]	43.1 %
<b>Our method</b>		<b>32.15 %</b>

## VI. CONCLUSION

A new multimodal annotation method is proposed. The method relies on the use of association rules mining and clustering; where the text and visual clusters obtained as output from the clustering algorithms are associated using the association rules mining. In order to evaluate the performance of the proposed method, two available and widely used datasets were considered to carry out the experiments. The results achieved with the proposed method outperforms all or most of the considered multimodal methods. The following improvements are considered as future work; the linking of tags to concepts using WordNet to find the semantic similarity between tags and concepts, the use of local visual features instead of global features. Improvement of text pre-processing and text features extraction in the training phase is necessary as well.

## ACKNOWLEDGMENT

We thank King Abdulaziz City for Science and Technology for funding this research (Grant No. 1-17-02-009-0003).

## REFERENCES

- [1] A. W. Smulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, December 2000.
- [2] Z. Mehmood, T. Mahmood, and M. A. Javid, "Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine," *Applied Intelligence*, vol. 48, no. 1, pp. 166–181, 2018.
- [3] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán, "Multiscale conditional random fields for image labeling," in *the Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, vol. 2. IEEE, June 2004, pp. 695–702.
- [4] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 163–168.
- [5] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *the Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. IEEE, July 2004, pp. 1002–1009.
- [6] A. Znaidia, "Handling imperfections for multimodal image annotation," Ph.D. dissertation, Ecole Centrale Paris, 2014.
- [7] W. B. Li, J. Min, and G. J. Jones, "A text-based approach to the imageclef 2010 photo annotation task," September 2010.
- [8] B. Sigurbjörnsson and R. Van Zwol, "Flickr tag recommendation based on collective knowledge," in *the Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 327–336.
- [9] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Image annotation refinement using random walk with restarts," in *the Proceedings of the 14th ACM international conference on Multimedia*. ACM, 2006, pp. 647–650.
- [10] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 119–126.
- [11] S.-B. Chan, H. Yamana, D.-D. Le, and S. Satoh, "Image annotation fusing content-based and tag-based technique using support vector machine and vector space model," in *the Proceedings of the 10th International Conference on Signal-Image Technology and Internet-Based Systems*. IEEE, April 2014, pp. 272–276.

- [12] Y. Zhang, S. Bres, and L. Chen, "Semantic bag-of-words models for visual concept detection and annotation," in *the Proceedings of the 8th International Conference on Signal Image Technology and Internet Based Systems*. IEEE, January 2012, pp. 289–295.
- [13] C.-m. Huang, Y.-S. Lee, C.-Y. Lin, and C.-Y. Chen, "Using LSA and Association Rules to Enhance Web Image Annotation," in *the Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 04 2011, p. 1.
- [14] S. Nowak, K. Nagel, and J. Liebetrau, "The imageclef2011 photo annotation and concept-based retrieval tasks," in *CLEF (Notebook Papers/Labs/Workshop)*, 2011, pp. 1–25.
- [15] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *the Proceedings of the European conference on computer vision*, vol. 2353. Springer, April 2002, pp. 97–112.
- [16] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134.
- [17] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *the Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management*. Citeseer, 1999, pp. 1–9.
- [18] M. Bakalem, N. Benblidia, and S. Ait-Aoudia, "A novel image auto-annotation based on blobs annotation," in *Image Processing and Communications Challenges 3*. Springer, 2011, vol. 102, pp. 113–122.
- [19] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, August 2015.
- [20] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *the Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.
- [21] N. Liu, E. Dellandréa, L. Chen, C. Zhu, Y. Zhang, C.-E. Bichot, S. Bres, and B. Tellez, "Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 493–512, May 2013.
- [22] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *European conference on computer vision*. Springer, 2008, pp. 316–329.
- [23] R. He, N. Xiong, L. T. Yang, and J. H. Park, "Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval," *Information Fusion*, vol. 12, no. 3, pp. 223–230, 2011.
- [24] M. Taïleb, S. Lamrous, and S. Touati, "Non-overlapping hierarchical index structure for similarity search," *International Journal of Computer Science*, vol. 3, no. 1, 2007.
- [25] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.
- [26] R. A. Alghamdi, M. Taïleb, and M. Ameen, "A new multimodal fusion method based on association rules mining for image retrieval," in *the Proceeding of 17th IEEE Mediterranean Electrotechnical Conference (MELECON)*. Beirut, Lebanon. IEEE, April 2014, pp. 493–499.
- [27] R. Agrawal, R. Srikant *et al.*, "Fast Algorithms for Mining Association Rules in Large Databases," in *the Proceeding of the 20th International Conference on Very Large Data Bases*, 1994, pp. 487–499.